DEFENCE R&D DÉFENSE

A Novel Method for Statistical Comparison of Geophysical Data by Multiple Instruments which have Differing Accuracies

T. Thayaparan, W.K. Hocking and S.J. Franke

Defence R&D Canada

TECHNICAL MEMORANDUM
DREO TM 2001-147
November 2001

National Defence Défense nationale

Canada

A Novel Method for Statistical Comparison of Geophysical Data by Multiple Instruments Which Have Differing Accuracies

T. Thayaparan
Defence Research Establishment Ottawa

W. K. Hocking
University of Western Ontario

S. J. Franke
University of Illinois

Defence Research Establishment Ottawa

Technical Memorandum

DREO TM 2001-147

November 2001

Author

Thayananthan Thayaparan

Approved by

Maria Rey
Head, Surface Radar Section

Approved for release by

Gordon Marwood
Chief Scientist, Defence Research Establishment Ottawa

Abstract

A novel correlative technique is introduced for comparison of measurements of similar quantities made using different techniques. This process involves a generalized least-squares fitting method which can be used to estimate the slope of the best-fitting straight line that results when two separate data sets that are expected to be linearly correlated are compared via scatter-plots. The different data types may be subject to different uncertainties in their measurements. The technique determines and graphs the relationship between the errors in each method and the slope of the line of best fit, assuming Gaussian statistics. We hope that this new technique will find application in a wide range of situations, e.g., ranging from radar and satellite measurements to any physical measurements. The technique can also be used to calibrate multi-sensor, radar and satellite, systems.

Résumé

Une technique corrélative innovatrice de comparaison de mesures de quantités similaires par des méthodes différentes est présentée. Ce processus fait intervenir une méthode généralisée d'ajustement par les moindres carrés pour l'estimation de la pente de la droite la mieux ajustée résultant de la comparaison par diagrammes de dispersion de deux jeux de données distincts dont on s'attend à ce qu'ils présentent une corrélation linéaire. Les mesures permettant d'obtenir différents types de données peuvent être sujettes à différentes incertitudes. La technique proposée permet de déterminer et de représenter graphiquement la relation entre les erreurs pour chaque méthode et la pente de la droite la mieux ajustée en supposant une répartition stochastique gaussienne. Nous espérons que cette nouvelle technique sera applicable à une gamme étendue de situations, p. ex. depuis les mesures effectuées au moyen du radar et de satellites jusqu'à toute mesure physique. Cette technique peut également servir à l'étalonnage de systèmes multicapteurs, radar et satellites.

Executive summary

One of the problems that frequently confronts an experimentalist is that of fitting a straight line to data, that is, the problem of determining the functional relationship between two variables. Least-squares techniques for examining best-fit lines to correlated data sets are well established, but in general are restricted to cases in which the errors in the two data sets being compared are known. The simplest case is that in which there is no error in one variable (usually plotted as the abscissa), and the error in the other variable is known. This procedure is described in almost all elementary books on statistics.

A more complex case is that in which both variables have errors, but both errors are well known. Examples which demonstrate how to deal with such cases have been shown in the literature. However, there are times when two data sets are compared in which the intrinsic measurement error of one or both variables are unknown. In this case, there is no algorithm available to determine the best-fit line. The situation can be further clouded if the different techniques measure similar, but not identical, parameters. Our purpose in this report is to consider optimal ways to compare such data.

A novel correlative technique is introduced for comparison of measurements of similar quantities made using different techniques. This process involves a generalized least-squares fitting method which can be used to estimate the slope of the best-fitting straight line that results when two separate data sets which are expected to be linearly correlated are compared via scatter-plots. The different data types may be subject to different uncertainties in their measurements. The technique determines and graphs the relationship between the errors in each method and the slope of the line of best fit, assuming Gaussian statistics. We hope that this new method will find application in a wide range of situations, e.g., ranging from radar and satellite measurements to any physical measurements. The technique can also be used to calibrate multi-sensor, radar and satellite, systems.

T. Thayaparan, W. K. Hocking, S. J. Franke. 2001. A Novel Method for Statistical Comparison of Geophysical Data by Multiple Instruments Which Have Differing Accuracies. DREO TM 2001-147. Defence Research Establishment Ottawa.

Sommaire

L'ajustement d'une droite à des données, c'est-à-dire la détermination de la relation fonctionnelle entre deux variables, constitue l'un des problèmes auxquels sont fréquemment confrontés les expérimentateurs. Les méthodes d'examen par les moindres carrés des droites les mieux ajustées à des jeux de données sont couramment utilisées, mais en général d'application limitée aux cas dans lesquels les erreurs dans les deux jeux de données comparées sont connues. Le cas le plus simple est celui dans lequel une des variables ne comporte pas d'erreur (habituellement celle qui est représentée en abscisse) et l'erreur entachant l'autre variable est connue. Cette procédure est décrite dans presque tous les ouvrages élémentaires de statistique.

La situation est plus complexe lorsque des erreurs entachent les deux variables, mais que ces erreurs sont bien connues. Des exemples de la manière dont il convient de traiter ces situations ont été documentés. Cependant, dans certains cas, il faut comparer deux jeux de données pour lesquels l'erreur intrinsèque de mesure d'une ou des deux variables n'est pas connue. Dans cette situation, il n'existe aucun algorithme permettant de déterminer la droite de meilleur ajustement. Le problème devient encore plus épineux si des méthodes différentes sont appliquées pour mesurer des paramétres similaires mais non identiques. Le présent rapport a comme objet l'examen de méthodes optimales de comparaison de telles données.

Une technique corrélative innovatrice de comparaison de mesures de quantités similaires par des méthodes différentes est présentée. Ce processus fait intervenir une méthode généralisée d'ajustement par les moindres carrés pour l'estimation de la pente de la droite la mieux ajustée résultant de la comparaison par diagrammes de dispersion de deux jeux de données distincts dont on s'attend à ce qu'ils présentent une corrélation linéaire. Les mesures permettant d'obtenir différents types de données peuvent être sujettes â différentes incertitudes. La technique proposée permet de déterminer et de représenter graphiquement la relation entre les erreurs pour chaque méthode et la pente de la droite la mieux ajustée en supposant une répartition stochastique gaussienne. Nous espérons que cette nouvelle technique sera applicable à une gamme étendue de situations, p. ex. depuis les mesures effectuées au moyen du radar et de satellites jusqu'à toute mesure physique. Cette technique peut également servir à l'étalonnage de systèmes multicapteurs, radar et satellites.

T. Thayaparan, W. K. Hocking, S. J. Franke. 2001. Une méthode innovatrice de comparaison statistique de jeux de données géophysiques recueillis au moyen d'instruments offrant différentes exactitudes. DREO TM 2001-147. Centre de Recherches pour la Défense Ottawa.

Table of contents

List of figures

1. Introduction

Least-squares techniques for examining best-fit lines to correlated data sets are well established, but in general are restricted to cases in which the errors in the two data sets being compared are known. The simplest case is that in which there is no error in one variable (usually plotted as the abscissa), and the error in the other variable is known. This procedure is described in almost all elementary books on statistics [1-5].

A more complex case is that in which both variables have errors, but both errors are well known. Examples which demonstrate how to deal with such cases have been shown by (amongst others) by [6-16].

However, there are times when two data sets are compared in which the intrinsic measurement error of one or both variables are unknown. In this case, there is no algorithm available to determine the best-fit line. The situation can be further clouded if the different techniques measure similar, but not identical, parameters. An example is shown in Figure 1, where two time-series of the eastward wind component in the atmosphere at an altitude of 85 km are shown. These data were recorded using two different techniques, one being a spaced antenna procedure applied to radiowave diffraction patterns produced by MF (medium frequency) radio scatter from ionospheric irregularities [17], and the other using interferometric applications applied to meteor trails formed in the Earth's atmosphere (e.g., see [18] and references there-in). Although each procedure ostensibly measures "winds" at 85 km, the two instruments measure over different areas and depths of the sky. As a result, there is a possibility that they measure different phenomena, especially if the winds show significant spatial variation. Our purpose in this paper is to consider optimal ways to compare such data. It should be emphasized here that the new proposed technique is not restricted to above example but can be applied in a wide range of situations, e.g., ranging from radar and satellite measurements to any physical measurements. The technique can also be used to calibrate multi-sensor, radar and satellite, systems.

This report begins by developing the required nomenclature, and a theoretical basis for the calculations. The basic theory for the ensuing procedures are then developed. In order to confirm the proposed theory, we simulate data using computer techniques, in which the intrinsic errors are pre-specified. We demonstrate the effects of employing at least one existing method of multi-error regression analysis [6] to these data sets, and point out the limitations of such procedures. Finally, we apply the proposed procedures to two sample data sets, including the points shown in Figure 1.

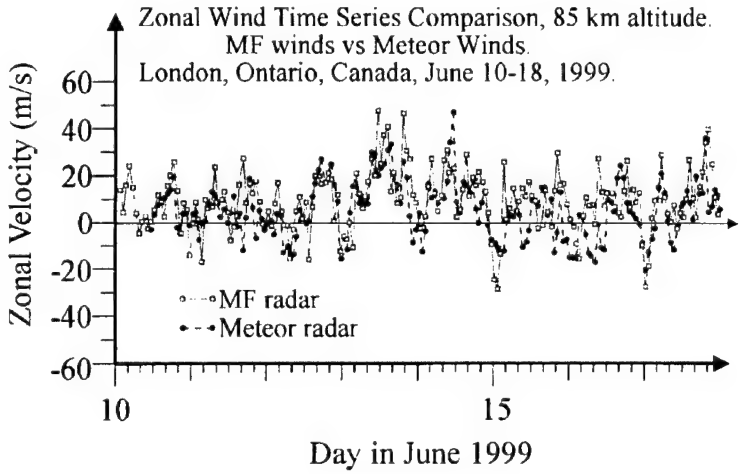

Figure 1: *An example of two time series of data obtained by two different models.*

2. Nomenclature and Theory

The goal in this section is to compare two data sets that are assumed to be linearly related, and to make meaningful statements about the relationship between them without having any a-priori knowledge about system errors. For the purposes of our theoretical developments, we will denote these data sets as $\{x_i\}$ and $\{y_i\}$ $(i = 1..N)$. The quantities $\{x_i\}$ are assumed to represent measurements of a randomly varying parameter, denoted here by $\{v_i\}$, and the quantities $\{y_i\}$ represent measurements of a parameter which is assumed to be linearly related to $\{v_i\}$. The data points $\{y_i\}$ are in fact assumed to have a constant, but unknown, gain relative to the $\{v_i\}$, which we will denoted as g_0. We will assume that each data set has a mean of zero: in practice, if this is not true, it is easily rendered so by removing the means from the data points. The model equations are therefore

(1) $$x_i = v_i + \delta x_i$$

(2) $$y_i = g_0 v_i + \delta y_i$$

where δx_i and δy_i are random (noise) components. We shall use the notation $N(\mu, \sigma^2)$ to denote the Gaussian probability density function with mean μ and variance σ^2. The following equations summarize our assumptions about the distribution of the measurement errors and the intrinsic associated parameters:

(3) $$v_i = N(0, \Sigma_v^2)$$

(4) $$\delta x_i = N(0, \sigma_x^2)$$

(5) $$\delta y_i = N(0, \sigma_y^2)$$

We also assume that v_i, δx_i and δy_i are mutually independent. Our purpose now is to determine an expression relating the variable g_0, Σ_v^2, σ_x^2 and σ_y^2.

We now square Eq. 1 and Eq. 2, and take ensemble averages (<>) to produce the following equations (remembering that we have assumed zero-mean quantities):

(6) $$< x_i^2 > = < v_i^2 > + < \delta x_i^2 > = \Sigma_v^2 + \sigma_x^2$$

(7) $$< y_i^2 > = g_0^2 < v_i^2 > + < \delta y_i^2 > = g_0^2 \Sigma_v^2 + \sigma_y^2$$

Furthermore, if we also multiply Eq. 1 by Eq. 2, and then take ensemble averages, we produced the additional equations

(8) $$< x_i y_i > = g_0 < v_i^2 > = g_0 \Sigma_v^2$$

Note that we have assumed that v_i, δx_i and δy_i are uncorrelated, so the averages of crossed terms are zero.

In what follows, the ensemble-averaged quantities involving the measured data are replaced with sample expectations. For example, $< x_i^2 >$ is replaced with $\xi_x^2 = (\Sigma_i x_i^2)/N$, $< y_i^2 >$ is replaced with $\xi_y^2 = (\Sigma_i y_i^2)/N$ and $< x_i y_i >$ is replaced with $\xi_{xy} = (\Sigma_i x_i y_i)/N$.

We also introduced two more variables in the following way. We recognize that the quantities $\{v_i\}$ are, in essence, the values of $\{x_i\}$ prior to the addition of the random (noise) component $\{\delta x_i\}$, and likewise the values $\{g_0 v_i\}$ are the values of $\{y_i\}$ prior to the addition of the noise component $\{\delta y_i\}$. We may consider these "non-noise" quantities as the "signal" embedded in our data. We henceforth use the definitions

(9) $$\Sigma_x^2 = \Sigma_v^2$$

(10) $$\Sigma_y^2 = g_0^2 \Sigma_v^2$$

where Σ_x^2 and Σ_y^2 represent the variances of the "signal" component of the $\{x_i\}$ and $\{y_i\}$ respectively. Then Eq. 6 and Eq. 7, combined with our definitions of ξ_x^2, ξ_y^2 and ξ_{xy}^2, produce the result

(11) $$\xi_x^2 = \Sigma_x^2 + \sigma_x^2$$

and

$$\xi_y^2 = \Sigma_y^2 + \sigma_y^2 \tag{12}$$

i.e., they tell us the somewhat obvious result that the overall variance of each data set is the sum of the variance associated with the signal and variance associated with the noise. Likewise we also have the result

$$\xi_{xy} = g_0 \Sigma_x^2 \tag{13}$$

We now turn to regression estimators. In many studies of regression, it is common to treat one variable as if it has no error, and assume all the error is associated with the second variable. For example, one might assume that $\sigma_x = 0$ in our above equations, while letting any error be associated with the variable y_i. Whilst this is untrue of our data, it makes a convenient starting point. If we take Eq. 11 and Eq. 13 and assume that $\sigma_x = 0$, and solve for g_0, then we produce an estimator for the slope which we will call $g_x{}'$. This quantity is given by

$$g_x{}' = \xi_{xy}/\xi_x^2 \tag{14}$$

This is exactly the same parameter as that deduced in any standard text book as the "slope of the least-squares best fit line" i.e., the regression of y on x [2,5]. The "best-fit" line has an equation

$$y_i = g_x{}' x_i + c_1, \tag{15}$$

where c_1 is a constant.

Likewise, by assuming that $\sigma_y = 0$, and solving for g_0 in Eq. 12 and Eq. 13, we produce the slope corresponding to the regression of x on y, which obeys the equation

$$x_i = (1/g_y{}')y_i + c_2, \tag{16}$$

viz.

(17) $$g_y{}' = \xi_y^2/\xi_{xy}.$$

Note that in the second case we have used an inverted form for the slope (i.e., it involves $1/g_y{}'$), so that we can write

(18) $$y_i = g_y{}' x_i + c_3,$$

Whilst neither of these estimators represent the true slope for our situation, they make excellent starting points for a derivation of a better estimator for g_0. In all real cases, $g_x{}'$ is less than or equal to g_0, and $g_y{}'$ is greater than or equal to g_0. The values $g_x{}'$ and $g_y{}'$ are also very easy to find, since many software packages produce these quantities as standard output (recognizing that the regression of x on y needs to have its slope inverted to be consistent with our definition of $g_y{}'$). We emphasize that, although these quantities were derived assuming either that $\sigma_x = 0$ or that $\sigma_y = 0$, we will not assume this in our more general analysis. These two slope estimators represent only a starting point for our calculations. For any data set, $g_x{}'$ and $g_y{}'$ are only estimators of population values g_x ($=< g_x{}' >$) and g_y ($=< g_y{}' >$), but as long as N is moderately large, we can take our estimates to be reasonable approximations of the population values, in the same manner that is usually applied to statistical studies.

Eq. 14 tells us that $\xi_{xy} = g_x \xi_x^2$, while Eq. 3 gives $\xi_{xy} = g_0 \Sigma_x^2$, so we may equate these two equations to give $\xi_x^2 = g_0 \Sigma_x^2$. Applying Eq. 11 and eliminating ξ_x^2 then gives

(19) $$\sigma_x/\Sigma_x = (g_0/g_x - 1)^{1/2}.$$

Likewise,

(20) $$\sigma_y/\Sigma_y = (g_y/g_0 - 1)^{1/2}.$$

Thus we see that the ratios of the "noise" variance to the "signal" variance are simple functions of g_0 and g_x for the x-series, and simple functions of g_0 and g_y for the y-series. Furthermore, we can use Eq. 11 and Eq. 12 to write

(21) $$\sigma_x = \xi_x(1 - g_x/g_0)^{1/2}.$$

Similarly,

$$\sigma_y = \xi_y(1 - g_0/g_y)^{1/2}. \tag{22}$$

Since ξ_x^2 and ξ_y^2 are just sample variances, and estimators for g_x and g_y can easily be determined for any data-set, the only unknown quantities in the above two equations are g_0, σ_x and σ_y. Since these three quantities are uniquely inter-related; specification of any one of them allows the determination of the other two. This is the key point of our procedure, and is a feature which we will employ throughout our future experimental examples.

3. Monte Carlo Simulations

In order to confirm our calculations above, we now turn to Monte Carlo modelling. Our simulations begin by developing a data-set of points $\{v_i{}'\}$ which have a Gaussian probability distribution with zero mean, as assumed in our previous theory. We assume that a Gaussian probability distribution is suitable for our data. The standard deviation of the set is written as Σ_v. Then, for each point $v_i{}'$, we calculate a point $v_{gi}{}' = g_0 v_i{}'$, so that the set $\{v_{gi}{}'\}$ has zero mean, and represents a re-scaling of the original set by g_0 time. Consistent with our description in the previous section, we consider these sets of points to be $\{x_i{}'\}$ and $\{y_i{}'\}$ values prior to the addition of random fluctuations, and therefore denote the standard deviation of the original set $\{v_i{}'\}$ as Σ_x, and the standard deviation for the quantities $\{v_{gi}{}'\}$ as $\Sigma_y = g_0\Sigma_v = g_0\Sigma_x$. These standard deviations essentially represent the "natural" spread of the (simulated) data set which we are generating. At this stage all the points $\{x_i{}', y_i{}'\}$ lie in a straight line if plotted in a "scatter plot" format. We now add a random component, by adding a vector $(\delta x_i, \delta y_i)$ to each point, where δx_i is a randomly selected value from a data-set with a Gaussian probability distribution of zero mean and standard deviation σ_x, and similarly δy_i is a randomly selected value from a separate data-set with a Gaussian probability distribution of zero mean and standard deviation σ_y. This produces scatter plots which look like those shown in Figure 2. We will denote this new set as $\{x_i, y_i\}$, consistent with the above theory.

Following this, we perform two regression analyses. First we treat the $\{x_i\}$ as if they have no errors, and produce a best-fit line with slope g_x (viz. the equation of this line is equation $y_i = g_x x_i + c_1$). This is the result of the regression of y on x. We then treat the $\{y_i\}$ points as if they have no error, and produce a best-fit line of slope g_y (i.e., we find the best-fit line satisfying $x_i = g_2 y_i + c_2$, but then convert it to the form $y_i = g_y x_i + c_3$, where $g_y = 1/g_2$). Examples are shown in Figure 2. It should especially be noted that the regression of y on x produces a slope which is less than the true slope, and likewise the regression of x on y also produces an underestimate of the true slope i.e., $g_x < g_0$ and $g_2 < g_0$. In the second case $g_2 < g_0$ means that $g_y > g_0$. In addition to performing these simulations, we also applied some existing dual-error algorithms to our data. In particular, we have applied the procedure of [6], using the known pre-assigned errors for each data-set.

We have performed the above calculation on many thousands of sets of points, and have determined that values of g_x and g_y for different combinations of Σ_x, Σ_y, σ_x, and σ_y. Figure 3 summarizes our results. It shows not only the average values of g_x and g_y for many different realizations for the case $g_0 = 1$ and pre-specified values of σ_x and σ_y (Figures 3a and 3c), but also the standard deviations (i.e., spread) in each estimate of g_x and g_y (right hand graphs). The results of the fitting procedures of [6] are also shown as Figures 3e and 3f. Figures 3d and 3f serve to demonstrate that [6]'s procedure works well, in that the slopes are always close to unity, but the method did require prior knowledge about the errors associated with each of the abscissa and the ordinate. If these errors are not known, the method cannot be applied, and our intent here is to deal

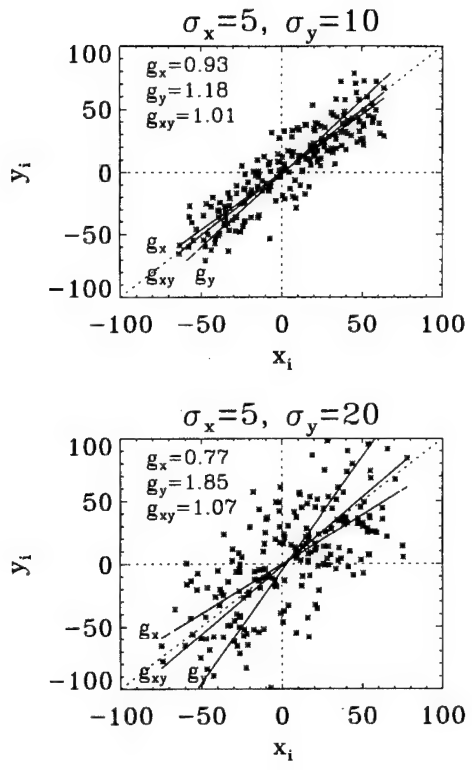

Figure 2: *Sample simulated scatter plots. In each case, we have taken* $\Sigma_x = \Sigma_y = 40$*, and used values for* σ_x *and* σ_y *as shown. The slopes* g_x *and* g_y*, which were determined by standard least-squares regression analysis, are indicated on the figure. The slope* g_{xy} *is the result of a fit due to York [6], which required prior knowledge of the values for* σ_x *and* σ_y*.*

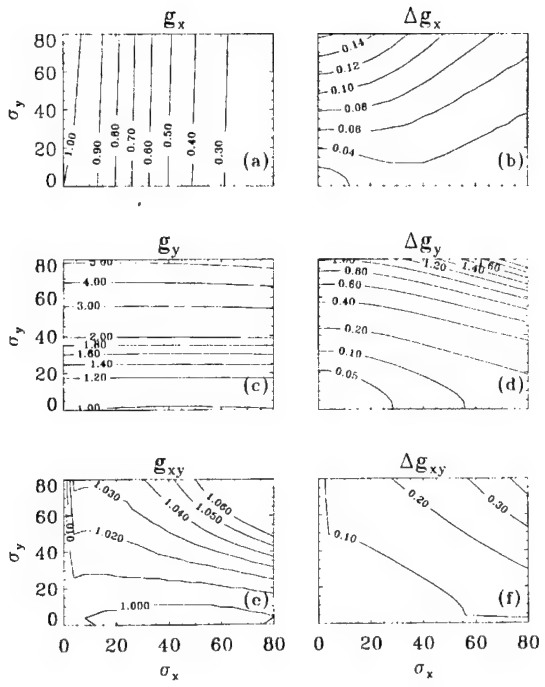

Figure 3: *Mean values of (a) g_x and (c) g_y calculated from our simulations, as functions of σ_x and σ_y, where in each case the means were determined from several hundred realizations (each realization being like the example shown in Figure 2, but each time using different sets of random numbers). We have used $\Sigma_x = \Sigma_y = 40$. In all cases the abscissa is σ_x and the ordinate is σ_y, and the values range from 0 to 80. Figures (b) and (d) show the standard deviations determined during these simulations, where (b) shows the standard deviations corresponding to (a), and (d) shows the standard deviations for (c). Figure (e) shows the results of calculations of the slope of the best-fit line using the procedure of York [6], where the known errors have been used in the calculations, and (f) shows the associated standard deviations for this procedure.*

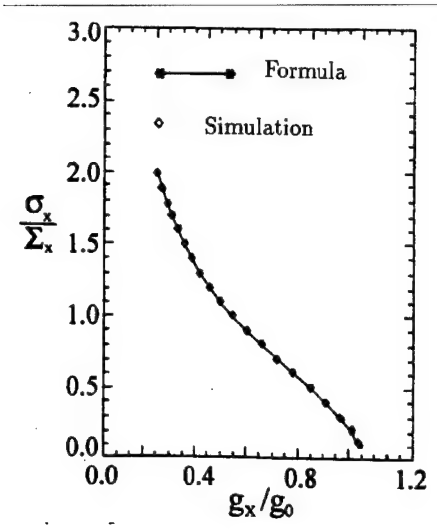

Figure 4: *Comparison of our calculated values of σ_x/Σ_x as a function of g_x/g_0, compared to our theoretical determinations (Eq 19).*

with situations with unknown errors. We shall therefore pursue the procedures of [6] no further, and will turn to a deeper discussion of Figures 3a, 3b, 3c and 3d.

With regard to Figures 3a and 3c, it should be noticed in particular that the contours are vertical in the first case, indicating that g_x is a function only of σ_x for a specified value of g_0, Σ_x and Σ_y, (which is consistent with Eq. 19), and the contours are horizontal for the second case, indicating that g_y is independent of σ_y (consistent with Eq. 20). We have also been able to show that a graph of σ_x/Σ_x as a function of g_x/g_0 has a form like that shown in Figure 4, irrespective of g_0, Σ_x and σ_x. Similarly we may determine graphs of σ_y/Σ_y as a function of g_0/g_y. As shown in Figure 4, our simulations match very closely Eq. 19 and Eq. 20. Figure 4, and Eq. 19 and Eq. 20 are the key to our calculations performed in this paper. We now turn to the application of these formulae in real situations.

To illustrated our applications to experimental measurements, we turn to Figure 5, which shows a scatter plot for an extended data series which includes the points in Figure 1. The data include both meridional and zonal components of the wind measured by co-located meteor and MF (Medium Frequency) radars at London, Ontario, Canada (see [18] for specific details about these innstruments). The graph in Figure 1 shows only the zonal component for a short period of 8 days, but our comparison uses a super-set of these data, covering the entire period from June 1 to July 31, 1999 (with some data gaps during which the radars were used for other purposes). For the present we will not concern ourselves with the physical meaning of the data, but will simply regard them as representative of any such correlative study between two such data sets. However, we do point out that the winds at 85 km altitude

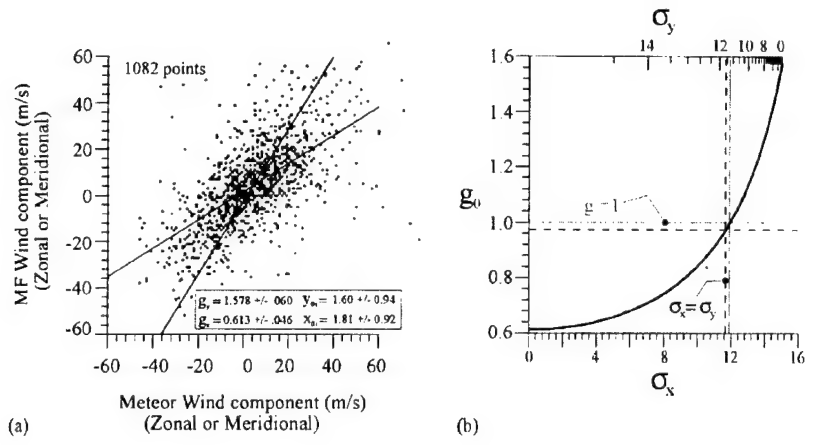

Figure 5: *Results of applying the fitting procedure to the data like those shown in fig. 1. The left-hand graph (a) shows the scatter plot of MF wind components vs. meteor wind components for an altitude of 85 km for the period of June and July, 1999, using co-located radars near London Ontario, Canada. Relevant parameters (offsets, g_x and g_y) are also shown on the figure. The quantities x_{0i} and y_{0i} refer to axis intercepts, as described in the text. Errors are quoted at the two-sigma level (approximately 95% significance). Fig (b) shows the relationship between g_0, σ_x, and σ_y, as described in the text.*

in the upper atmosphere are highly variable, both temporally and spatially, and comprise multiple large-amplitude waves. Vertical excursions as large as 5-6 km in a few hours are not uncommon. Hence the reader should not be surprised by an apparent low correlation between our two data sets; the higher correlations which are typical of (for example) tropospheric data are simply not possible in upper atmospheric cases.

When presented with real data, we perform the following calculations. We first find the standard deviation of each dataset $\{x_i\}$ and $\{y_i\}$, which we will denote as ξ_x and ξ_y. We then also determine the slopes of the regression of y on x, to give g_x, and the regression of x on y, thereby obtaining g_y (remembering that the slope of the regression of x on y must be inverted to give g_y). We also determine intercepts of these best-fit lines with the axes, which are denoted as x_{0i} and y_{0i} in this (and subsequent) figures. The value x_{0i} refers to the intercept of the line of slope g_x with the ordinate, and the value y_{0i} refers to the intercept of the line of slope g_y with the abscissa. Following this, we assume various values for σ_x, starting at 0.0 and stepping up in small steps. For each value of σ_x, we calculate a value for Σ_x through the relation $\Sigma_x = (\xi_x^2 - \sigma_x^2)^{1/2}$ (from Eq. 11). We can then deduce the ratio σ_x/Σ_x, and from the graph in Figure 3, or from Eq. 19, we therefore deduce g_x/g_0. Knowing g_x, we can therefore deduce g_0 for this particular choice of σ_x. Following this, we can deduce Σ_y because $\Sigma_y = g_0\Sigma_x$. Finally, we may deduce $\sigma_y = (\xi_y^2 - \Sigma_y^2)^{1/2}$. We may therefore plot the relationship between g_0, σ_x and σ_y as shown in Figure 5b. Note that the upper axis (σ_y) requires a

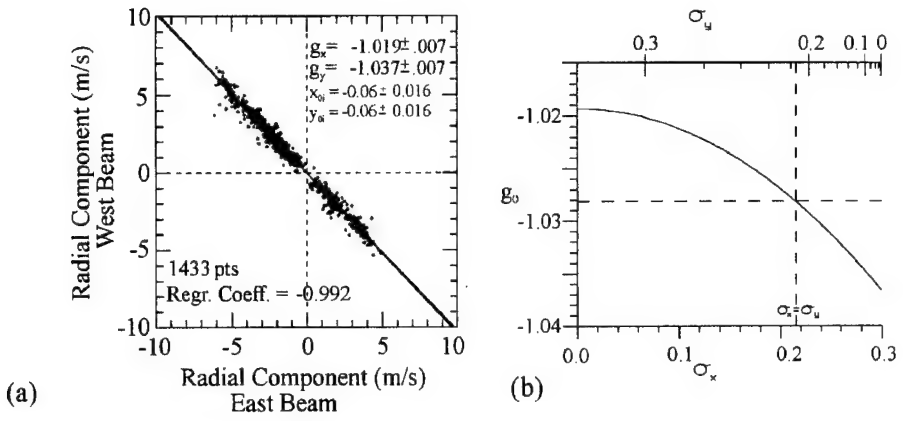

Figure 6: *As for Figure 5, but using data from a tropospheric wind-profiler (see text).*

different scaling for each new data-set for which this procedure is applied.

This procedure therefore explains how the secondary graphs in Figure 5 were produced. Extra information is required in order to determine the "correct" combination of g_0, σ_x and σ_y. (Thayaparan and Hocking [19] attempted to use an additional graph like that shown in Figure 4, but relating σ_y/Σ_y to g_y/g_0, to try and uniquely determine g_0, σ_x and σ_y, and initial trials suggested some success. However, it subsequently evolved that the information available with this process was redundant, and a unique determination was impossible without further information. Therefore the reader should be aware that Figure 4 of [19], is incorrect, as are the developments following Eq. 9 of that work.)

In all of our scatter plots we will henceforth include graphs like that shown in Figure 5b. This graph encompass many of the features of other fitting algorithms. For example, if the user has sufficient justification to assume $\sigma_x = \sigma_y$, then g_0 can be read from the graph directly. If a particular value of σ_x (or for that matter σ_y) is known, then again g_0 can be found uniquely. If it is known that g_0 is 1.0 (for example), then σ_x and σ_y can be immediately determined. The so-called "OLB" (Ordinary Least-squares Bisector") method (in which the sum of the squares of the perpendicular distances between the best-fit line and the points is minimized) is another subset of this curve, and corresponds to the case where the ratio of σ_x to σ_y is assumed to be proportional to g_0. In Figure 5b, the cases where $\sigma_x = \sigma_y$ and $g_0 = 1.0$ are almost coincident, but this need not always be so. Thus this graph encompasses many of the features available with other least squares algorithms, but gives a much clearer picture of the relationship between these various variables. Most other algorithms discussed in the current literature produce subsets of the more general $(g_0, \sigma_x, \sigma_y)$ combinations shown in figures like Figure 5b.

Figure 6 shows another example of application of this process. In this case, the data were taken using a tropospheric radar, so the scatter is less severe. This is true because the wind fields in the troposphere tend to be much less chaotic than those in the upper atmosphere. The data refer to radial Doppler velocities measured on two beams pointing at 10.9 degrees off-vertical in opposite directions. Again, the various possible combinations of g_0, σ_x and σ_y are defined by Figure 6b. In this case, because the beams are essentially identical, it would be sensible to surmise that $\sigma_x = \sigma_y$, so that we may determine that $g_0 = -1.028$. This therefore tells us that the effective beams in each case are not at the same zenithal angle, but differ by about 18 minutes of arc. The exact reason for this is not important for this paper, but our statistical procedure has been able to determine this asymmetry.

4. The Meaning of σ_x and σ_y

Finally, we need to consider what σ_x and σ_y really represent. If we are using two instruments to measure the same quantity, then these two variables tell us the measurement error in each technique. However, in many situations in geophysics we cannot measure exactly the same phenomenon. The comparison between the MF and meteor winds which we considered in Figure 5 is one such example. Whilst it would be desirable for each instrument to measure in exactly the same region of space, over the same interval of time, this is simply not possible. Thus the values of the quantities σ_x and σ_y reflect not only the intrinsic measurement errors of each instrument, but also contain information about how "different" the two techniques are. In essence, the least-squares fitting procedure indirectly assumes that there is a "compromise" measurement technique that is neither the MF radar nor the meteor method, and assumes that each instrument is trying to achieve this compromise. For example, as a crude example, the MF radar measures in a volume of space which is typically 10-20 km across, whilst the meteor radar measures over a region of space of perhaps a 100 km or more in diameter. This "compromise" measurement might represent an instrument which could measure in an intermediate volume of space - perhaps with a width of fifty km or so and a depth of 3-4 km. Of course this grossly oversimplifies the true situation, but illustrates the concept. In reality, we can never actually state what this "compromise technique" is, or even achieve it. However, we can say that σ_x and σ_y contain information about both the intrinsic measurement errors of each technique and also the level of departure of each of the MF and meteor measurements from this (unknown) compromise. Thus any values of σ_x and σ_y need to be interpreted both in terms of measurement error and also the natural spatial and temporal variability of the medium in which the measurements are being undertaken, as well as the design of the experiment being used to perform the measurements.

An alternative way to consider the situation is to consider that there are two aspects to our measurements - a component which is fully correlated between the two techniques, and an uncorrelated component. With this viewpoint, the fully correlated component represents the common features of the measurements, while the terms σ_x and σ_y represent the departures of the measurements from this correlated component, and therefore are a measure of the contribution of each technique to the uncorrelated component.

It is also apparent that at some times of observation in Figure 1 (e.g., the morning of June 16), the two different measurements are very different - in some cases, even opposite in sign. Close investigation of such events would probably reveal some sort of large spatial variability, such as a large localized buoyancy wave in the field of view, which might be seen by one technique but not the other. Of course such occurrences must happen, and deserve further scrunity. However, they are not the subject of our treatment here, since we wish only to examine statistical differences between the two methods - anomalies like these points are simply considered as part of the natural course of statistical fluctuations.

Despite the fact that the techniques being compared may often have some degree of non-commonality, useful information is forthcoming from these graphs. First, they place limits on the allowable values of g_0. For example, if $g_0 = 1$ is not an allowable solution, we can states that the two methods are different (as seen, for example, in relation to Figure 6). Secondly, the values of σ_x and σ_y place upper limits on the intrinsic instrumental errors. Thirdly, the relation between g_0, σ_x and σ_y is clearly specified by graphs like Figure 5b. The procedure therefore can be quite instructive in relation to comparisons between different measuring techniques. We have concentrated on cases where the expected measurements are in the ratio 1:1, or 1:-1, but the procedure is quite general. It can be used in many other cases where it is known that 2 measurements are proportional, but may not be in the ratio 1:1. In this case, the information about limits on the optimum values for the "gain" of one method over the other (i.e. the value of g_0) can be determined.

5. The Interpretation of Correlation Coefficient

The correlation coefficient, ρ, for data sets $\{x_i\}$ and $\{y_i\}$ is:

$$(23)\quad \rho = \frac{< x_i >< y_i >}{\sqrt{< x_i^2 >< y_i^2 >}} = \frac{1}{\sqrt{(1+\frac{\sigma_x^2}{\Sigma_x^2})(1+\frac{\sigma_y^2}{\Sigma_y^2})}} = \frac{1}{\sqrt{(1+\frac{\sigma_x^2}{\Sigma_v^2})(1+\frac{\sigma_y^2}{g_0^2\Sigma_v^2})}}$$

Notice that the correlation depends on all of the parameters σ_x, σ_y, g_0, and Σ_v. Now, a quantitative test of whether two measurements "agree" or "disagree" (e.g., a statistical hypothesis test) involves calculation of ρ from the data (using sample means in the numerator and denominator) and a comparison of this number with a prediction based on known values of σ_x, σ_y, g_0, and Σ_v. Thus quantitative interpretation of the numerical value of ρ requires knowledge of all of these parameters. Without knowledge of these parameters, it is impossible to judge whether a numerically small correlation results because the instruments are measuring different physical quantities or because the instruments are measuring the same quantity, but with observation errors that exceed the variability of the measured quantity (i.e., $\sigma_x > \Sigma_x$ and/or $\sigma_y > \Sigma_y$).

For example, suppose the observation errors, modelled by σ_x and σ_y, are equal to the geophysical variabilities, Σ_x and Σ_y, i.e., suppose $\sigma_x = \Sigma_x$ and $\sigma_x = \Sigma_x$. In this case $\rho = 0.5$. Thus, we can conclude that when observations errors equal or exceed the underlying geophysical variability, measured correlation coefficients will be smaller than 0.5. This does not necessarily indicate that one or the other measurement is "correct" in any sense. It may simply indicate that one or both of the data sets is "noisy".

For further illustration, consider the following "thought problem" wherein the geophysical variability is zero. Suppose that the actual velocity in some region of the mesosphere is exactly zero. Consider two instruments which make independent observation errors and let those instruments repeatedly probe the region to produce a time series of velocity estimates. Let the observation errors σ_x and σ_y be arbitrarily small (e.g., suppose they are on the order of 1 cm/s or less). Then each instrument will produce a time series consisting of small velocity estimates, on the order of ± 1 cm/s. The corrlation between the two time series will be zero ! And yet each instrument is making a high precision measurement of the true velocity, because the mean value will be removed from the measurements when computing the correlation coefficient. Obviously, this is an extreme example, but it illustrates that a small correlation coefficient is essentially meaningless without additional, and often unavailable, side information about the geophysical variability of the underlying quantity that is being measured.

In summary, high correlation always indicates agreement between two techniques, but low correlation does not necessarily indicate "poor agreement" between two

techniques. Thus, correlation coefficients can not be used to show that two techniques disagree (and hence, to show that one of the techniques is "wrong") without a careful assessment of the relative magnitudes of observation errors and geophysical variability. This often requires more information than is available.

6. Conclusions

A procedure has been described which permits meaningful comparison of different measurements of correlated quantities, using a scatter-diagram approach. It is emphasized that linear least-squares regression techniques always result in an under-estimate of the true slope, and the need to recognize that errors exist in both the abscissa points and the ordinate points must be recognized. The inter-relationship between the intrinsic system errors and the optimum slope is derived, and presented in a graphical manner. Useful limits can be placed on the errors and the optimum slope. We hope that this new method will find application in a wide range of situations, e.g., ranging from radar and satellite measurements to any physical measurements. The technique can also be used to calibrate multi-sensor, radar and satellite, systems.

References

1. Young, H. D. (1962). Statistical Treatment of Experimental Data, *McGraw-Hill,* New York.

2. Bevington, P. R. (1969). Data Feduction and Error Analysis for the Physical Sciences, *McGraw-Hill,* New York.

3. Daniel, C. and Wood, F. S. (1971). Fitting equations to data, *Wiley-Interscience,* New York.

4. Brandt, S. (1976). Statistical and computational methods in data analysis, *Elsevier,* New York.

5. Taylor, J. R. (1982). An Introduction to Error Analysis, *Mill Valley,* California, U.S.A..

6. York, D. (1966). Least-squares Fitting of a Straight Line, *Can. J. Phys.,* 44, 1079-1086.

7. Barker, D. R. (1974). Simple Method for Fitting Data when Both Variables Have Uncertainties, *Am. J. Phys.*, 42(3), 224-227.

8. Orear, J. (1982). Least-squares when Both Variables have Uncertainties, *Am. J. Phys.,* 50(10), 912-916.

9. Lybanon, M. (1984). A Better Least-squares Method when Both Variables Have Uncertainties, *Am. J. Phys.,* 52(1), 22-26.

10. Miller, B. P. and Dunn, H. E. (1988). Orthogonal Least-squares Line Fit with Variable Scaling, *Computers in Physics,* July/August.

11. Reed, B. C. (1989). Linear Least-squares Fits with Errors in Both Coordinates, *Am. J. Phys.,* 57(7), 642-646.

12. Reed, B. C. (1992). Linear Least-squares Fits with Errors in Both Coordinates: Comments on Parameter Variances, *Am. J. Phys.,* 60(1), 59-62.

13. Babu, G.J., and Fiegelson, E. D. (1992). Analytical and Monte Carlo Comparisons of Six Different Linear Least-squares Fits, *Communications in Statistics - Simulation and Computation*, 21(2), 533-549.

14. Feigelsen, E.D. and Babu, G. J. (1992). Linear Regression in Astronomy, II, *Astrophysical Journal,* 397, 55-67.

15. Macdonald, J. R., and Thompson, J. (1992). Least-squares Fitting when Both Variables Contain Errors; Pitfalls and Possibilities, *American J. Physics*, 60(1), 66-73, 1992.

16. Jolivette, P. L. (1993). Least-squares Fits when there are Errors in X, *Computers in Physics,* 7, 208-212.

17. Briggs, B.H. (1984). The Analysis of Spaced Sensor Records by Correlation Techniques, *Handbook for MAP,* Ground based techniques, vol. 13, pp 166-186, Univ. of Illinois, Urbana.

18. Hocking, W.K., and Thayaparan, T. (1997). Simultaneous and Co-located Observation of Winds and Tides by MF and Meteor Radars Over London, Canada, (43° N, 81° W) During 1994-1996, *Radio Sci.,* 32, 833-865.

19. Thayaparan, T. and Hocking, W. K. (1998). A Least-squares Straight-line Fitting Algorithm with Automatic Error Determination, *Technical Note 98-004, Defense Research Establishment Ottawa,* June.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Establishment sponsoring a contractor's report, or tasking agency, are entered in section 8.)

Defence Research Establishment Ottawa
Department of National Defence
Ottawa, Ontario, Canada K1A 0Z4

2. SECURITY CLASSIFICATION (overall security classification of the document, including special warning terms if applicable)

UNCLASSIFIED

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S,C or U) in parentheses after the title.)

A Novel Method for Statistical Comparison of Geophysical Data by Multiple Instruments Which Have Differing Accuracies (U)

4. AUTHORS (Last name, first name, middle initial)

Thayaparan, Thayananthan, Hocking, Wayne and Franke, Steve

5. DATE OF PUBLICATION (month and year of publication of document)	6a. NO. OF PAGES (total containing information. Include Annexes, Appendices, etc.)	6b. NO. OF REFS (total cited in document)
November 2001	21	19

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

DREO TECHNICAL MEMORANDUM

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include the address.)

Defence Research Establishment Ottawa
Department of National Defence
Ottawa, Ontario, Canada K1A 0Z4

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant)	9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written)
05AB11	

10a. ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NOS. (Any other numbers which may be assigned this document either by the originator or by the sponsor)
DREO TECHNICAL MEMORANDUM	

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

(X) Unlimited distribution
( ) Distribution limited to defence departments and defence contractors; further distribution only as approved
( ) Distribution limited to defence departments and Canadian defence contractors; further distribution only as approved
( ) Distribution limited to government departments and agencies; further distribution only as approved
( ) Distribution limited to defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in 11) is possible, a wider announcement audience may be selected.)

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual).

(U) A novel correlative technique is introduced for comparison of measurements of similar quantities made using different techniques. This process involves a generalized least-squares fitting method which can be used to estimate the slope of the best-fitting straight line that results when two separate data sets that are expected to be linearly correlated are compared via scatter-plots. The different data types may be subject to different uncertainties in their measurements. The technique determines and graphs the relationship between the errors in each method and the slope of the line of best fit, assuming Gaussian statistics. We hope that this new technique will find application in a wide range of situations, e.g., ranging from radar and satellite measurements to any physical measurements. The technique can also be used to calibrate multi-sensor, radar and satellite, systems.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Least-Squares Fitting
Error Analysis
Linear Regression
Best-fit Line
Correlation Coefficient
Data Analysis
Normal distribution
Uncertainties
Measurement Errors
Standard deviation
Variance
Slope
Gradient